Computer Assisted
Language Instruction
Consortium

CALICO JOURNAL (ONLINE) ISSN 2056–9017

Article

Investigating the Application of Automated Writing Evaluation to Chinese Undergraduate English Majors: A Case Study of *WriteToLearn*

Sha Liu and Antony John Kunnan

Abstract

This study investigated the application of WriteToLearn *on Chinese undergraduate English majors' essays in terms of its scoring ability and the accuracy of its error feedback. Participants were 163 second-year English majors from a university located in Sichuan province who wrote 326 essays from two writing prompts. Each paper was marked by four human raters as well as* WriteToLearn. *Many-facet Rasch measurement (MFRM) was conducted to calibrate* WriteToLearn*'s rating performance in scoring the whole set of essays against those of four trained human raters. In addition, the accuracy of* WriteToLearn*'s feedback on 60 randomly selected essays was compared with the feedback provided by human raters. The two main findings related to scoring were that: (1)* WriteToLearn *was more consistent but highly stringent when compared to the four trained human raters in scoring essays; and (2)* WriteToLearn *failed to score seven essays. In terms of error feedback,* WriteToLearn *had an overall precision and recall of 49% and 18.7% respectively. These figures did not meet the minimum threshold of 90% precision (set by Burstein, Chodorow, and Leacock, 2003) for it to be considered a reliable error detecting tool. Furthermore, it had difficulty in identifying errors made by Chinese undergraduate English majors in the use of articles, prepositions, word choice and expression.*

Keywords: Accuracy of error feedback; automated writing evaluation; Chinese undergraduate English majors; scoring ability; *WriteToLearn*

Affiliation

School of Foreign Languages, China West Normal University, Nanchong City, Sichuan Province, 637000, China.
email: liusha.sally@yahoo.com (corresponding author)

CALICO JOURNAL VOL 33.1 2016 71–91
doi : 10.1558/cj.v33i1.26380

Introduction

The last few decades have seen the development of Automated Essay Soring (AES) systems for instructional applications including *e-rater* by Educational Testing Service (ETS), *Intelligent Essay Assessor (IEA)* by Pearson Education Inc., *Intellimetric* by Vantage Learning, *Bookette* by CTB/McGraw-Hill, and *LightSIDE* by Turnitin LLC. with corresponding instructional systems: *Criterion*, *WriteToLearn*, *My Access, Writing Roadmap,* and *Turnitin*. In addition to the primary function of essay scoring, these systems are trained to provide automated feedback on errors in students' essays (i.e., error feedback) and offer suggestions on discourse elements such as organization and development of ideas (i.e., discourse feedback). With this combination of automated scoring and feedback, these instructional systems are now referred to as automated writing evaluation (AWE) systems (McNamara, Crossley, Roscoe, Allen, & Dai, 2015). Particularly in the last decade, these AWE systems have been gradually marketed in classroom settings with English language learners (ELLs) although they were initially designed for native speaker-writers of English in the US (Warschauer & Ware, 2006). As there is insufficient research evidence to support the instructional application in the ELL setting, questions have been raised regarding AWE systems' ability to score and detect errors in essays written by ELLs (Weigle, 2013a).

This study investigated the application of *WriteToLearn,* the AWE system that has now been marketed to schools and colleges in China. The aim of the study was to systematically evaluate *WriteToLearn*'s performance in scoring essays written by Chinese undergraduate English majors and its accuracy in providing error feedback when compared to trained human raters.

AWE systems' scoring ability

An extensive review showed that the vast majority of studies of automated scoring have been conducted or funded by AWE systems developers and mainly based on data from native English speaker-writers in large-scale writing assessments in the US. Numerous studies were conducted to evaluate AWE systems' scoring ability, including *e-rater* (Attali & Burstein, 2005; Burstein *et al*., 1998; Burstein, 2003), *Intellimetric* (Vantage Learning, 2003a, 2003b, 2006), and *IEA* (Foltz, Laham, & Landauer, 1999; Landauer, Laham, & Foltz, 2003; Foltz, Lochbaum, & Rosenstein, 2011; Foltz, Streeter, Lochbaum, & Landauer, 2013). Table 1 presents results from three AWE systems in terms of correlations and agreement indices between human rater scores and automated scores.

Table 1: Correspondence Indices between Human Rater Scores and Automated Scores

AEE systems	Correlation (*r*)	Agreement (%)	
		Exact	Adjacent
e-rater	0.79–0.87 (Burstein *et al*., 1998)	45–59 (Attali & Burstein, 2005)	87–97 (Burstein, 2003)
Intellimetric	0.93 (average) (Vantage Learning, 2006)	76 (average) (Vantage Learning, 2006)	99 (average) (Vantage Learning, 2006)
IEA	0.76–0.95 (Fotlz *et al*., 2011)	50–81 (Fotlz *et al*., 2011)	91–100 (Fotlz *et al*., 2011)

Note: Adjacent agreement refers to the agreement within an acceptable discrepancy threshold. For example, with a six-point rating scale, adjacent agreement refers to the agreement within one score point.

Collectively, the findings of these studies suggested that 'AES correlates well with human-rater behavior, may predict as well as humans, and possesses a high degree of construct validity' (Shermis & Burstein, 2003: xiii). Likewise, in a recent study that examined nine AWE systems' scoring ability, Shermis (2014) claimed that '[a]utomated essay scoring appears to have developed to the point where it can consistently replicate the resolved scores of human raters in high-stakes assessment' (p. 23).

Such findings, however, should be considered with 'a highly critical eye' (Warschauer & Ware, 2006: 163) for many reasons. First, research conducted by instructors/independent researchers in ELL contexts does not support the results reported by AWE developers and developer-sponsored researchers. Hoang (2011) showed a moderate correlation between *Intellimetric* scores and averaged scores assigned by two human raters ($r = 0.688$, $p < 0.05$) which is in contrast to the strong correlation between the human raters ($r = 0.783$, $p < 0.05$). Similarly, Li, Link, Ma, Yang, and Hegelheimer (2014) reported that *Criterion* scores for a narrative topic correlated moderately with those by human raters (Spearman's $\rho = 0.426$, $p < 0.05$) and there was no statistically significant correlation between *Criterion* scores for the argumentative topic and human rater scores. Instructors and independent researchers have also expressed concerns about the validity of automated scores because AWE systems fail to 'read' the essay content (Hoang, 2011; McGee, 2006; Powers, 2000). It has been asserted that *IEA* is able to capture the text's meaning in the same way as human readers do (Pearson Education Inc., 2010). McGee's (2006) study challenged this claim, showing that: (a) *IEA* failed to take into consideration the global arrangement, cohesion, and coherence in its evaluation of 'meaning', despite the fact that they are of great importance to the meaning construction of any written text; (b) *IEA* claimed to measure factual content but awarded a high score of five to an essay with inaccurate factual content; and (c) *IEA* assigned an extremely high overall score of seven to an essay which made no sense to human raters. Similarly, Hoang (2011) showed that *Intellimetric* was not able to detect off-topic and plagiarized essays that were identified by human raters.

Second, high agreements or reliability statistics of scores as demonstrated by relatively strong correlations with human rater scores do not equal validity (Attali, 2013). Although apparently straightforward and easy to understand, percent agreement (exact and adjacent agreement) could be deceptive: a relatively high agreement rate can be expected solely by chance if raters only use a few points on the rating scale (Bridgeman, 2013; Powers, 2000). The worst scenario, as persuasively argued by Bridgeman (2013), could be that if both the AWE system and a human rater happen to assign a certain band score (for example, three out of six) to every essay, the exact agreement rate would be 100%. However, the agreement rate in this case would be unreliable in terms of evaluation of students' writing competence and AWE system's scoring performance. It is this concern that has propelled researchers to use many-facet Rasch measurement (MFRM) as an alternative method to Classical True Score Theory in investigating the relationship between automated scores and human rated scores (Koskey & Shermis, 2013).

Third, the fact that automated scoring and human ratings produce similar results does not automatically mean that the ratings are perfect. The ratings need to be examined because it could also be the case that that the ratings from the two scoring methods were based on invalid features such as essay length which is not entirely relevant to high-quality writing (Landauer *et al.*, 2003) but well known as a strong predicator of automated scores (Perelman, 2014; Powers, Burstein, Chodorow, Fowels, & Kukich, 2002). Recently, in a study reported in Powers *et al.*, (2002), *e-rater* assigned the highest possible score of six to an essay in which several paragraphs were repeated 37 times; in contrast, two human raters awarded the lowest possible score of one to the essay.

AWE systems' error feedback accuracy

There has also been a lingering concern regarding AWE systems' accuracy in detecting errors committed by ELL writers. This concern is due to the fact that ELL student writers tend to use numerous non-native expressions and make errors that are not frequently found in native English speakers' writing (Leacock, Chodorow, Gamon, & Tetreault, 2010). To address this concern, a number of studies have been conducted to evaluate the accuracy of AWE systems' error feedback, either by AWE developers (Burstein *et al.*, 2003; Han, Chodorow, & Leacock, 2006; Tetreault & Chodorow, 2008a) or by independent researchers (Dikli & Bleyle, 2014; Hoang, 2011).

A review of these studies showed that the detection of ELL errors, especially word choice, preposition and article errors, remains to be a thorny issue for current AWE systems. For example, *Criterion*'s accuracy in detecting preposition errors shows that its precision and recall are reported to be 84% and 19% respectively (Tetreault & Chodorow, 2008a). This means that it can only detect 19% of the preposition errors that human annotators (such as teachers) would identify with 84% accuracy. In detecting article errors, *Criterion*'s precision and recall are 90% and 40%, respectively (Han *et al*., 2006). Dikli and Bleyle (2014) reported similar results: they found that preposition and article errors were largely misidentified by *Criterion* while word choice and word form errors were highly under-identified.

In an attempt to evaluate the accuracy of *My Access*'s error feedback, Hoang (2011) found that *My Access*, like *Criterion,* was unable to detect word choice and word form errors. In detecting preposition errors, the precision and recall were about 78% and 19% respectively. Although *My Access*' precision and recall in detecting article errors (about 97% and 25% respectively) were higher than those of *Criterion*, this finding cannot be generalized because of the small sample size (n = 15) and the use of a single human rater as the gold standard (Tetreault & Chodorow, 2008b). In summary, current AWE systems did not meet the minimum threshold of 90% precision (as set by Burstein *et al*., 2003). for them to be reliable error detecting tools.

Research questions

Based on the above review of the literature and the limitations and research gaps that were revealed, two research questions were articulated for the study:

1. How does *WriteToLearn* perform in scoring Chinese undergraduate English majors' essays compared to scoring by trained human raters?
2. How accurate is *WriteToLearn*'s feedback to Chinese undergraduate English majors compared to trained human raters?

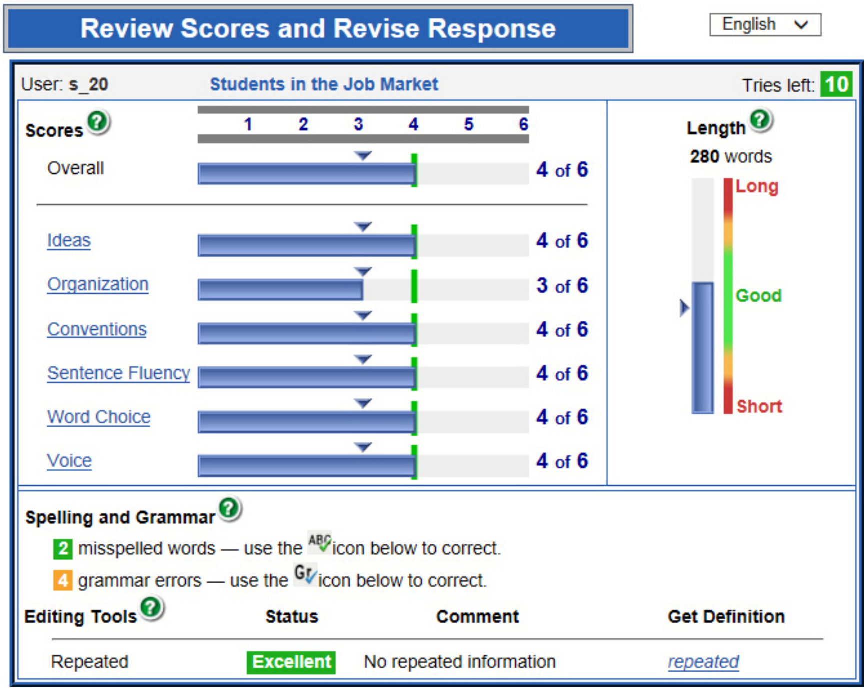
Methodology

Participants

The study participants were 163 undergraduate EFL learners (male = 9; 5.5%; female = 154; 94.5%) enrolled in English Education in a major university in the Sichuan province of the People's Republic of China. Their age ranges between 19 and 23 years and overall, they had been learning English for approximately 9.5 years at the time of the study.

WriteToLearn

We used *WriteToLearn* (V9.0) developed by Pearson Education Inc. (http://www.writetolearn.net/). *WriteToLearn* contains an assignments library of over 650 writing prompts for students in grades 3–12 in various genres including expository, descriptive, narrative, persuasive, and argument. The system also allows users to create their own writing prompts. When scoring essays, *WriteToLearn* generates a set of scores (both analytical and holistic). Figure 1 illustrates the overall score and six trait scores awarded for a sample student essay. By clicking on individual traits, students can receive detailed explanations of how to improve those particular aspects of writing.



Review Scores and Revise Response

English

User: s_20 Students in the Job Market Tries left: 10

Scores	
Overall	4 of 6
Ideas	4 of 6
Organization	3 of 6
Conventions	4 of 6
Sentence Fluency	4 of 6
Word Choice	4 of 6
Voice	4 of 6

Length

280 words

Spelling and Grammar

2 misspelled words — use the ABC icon below to correct.

4 grammar errors — use the Gr icon below to correct.

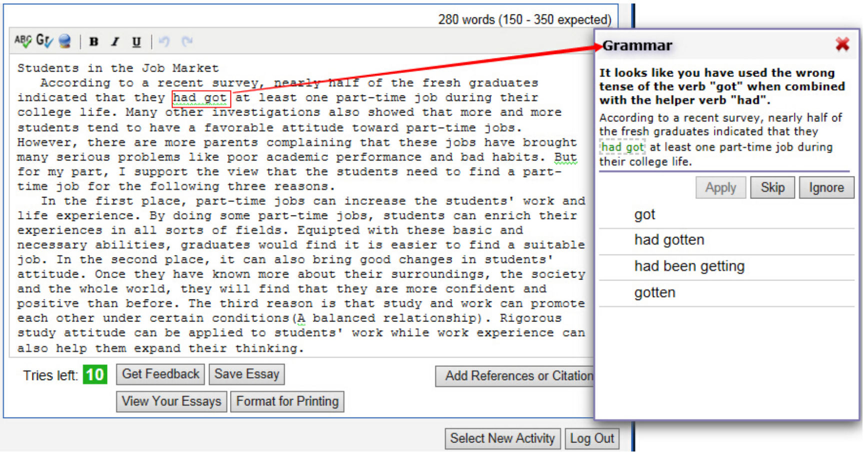
Editing Tools	Status	Comment	Get Definition
Repeated	Excellent	No repeated information	repeated

Figure 1: Screenshot of *WriteToLearn*'s scoreboard

When detecting errors, *WriteToLearn* did not tag errors in students' essays; instead, it briefly explained the problems and offered suggestions to correct the errors. As shown in Figure 2, *WriteToLearn* provided an explanation like 'it looks like you have used the wrong tense of the verb got when combined with the helper verb had' instead of marking the error as 'verb tense error'.

Raters and rater training

Four human raters were recruited for rating the essays. Table 2 summarizes the demographic information of the raters. Two raters were native English speakers and the other two were highly proficient ELLs whose first language was Chinese. All raters had past experience in teaching English at tertiary level institutions and marking essays.



Students in the Job Market

According to a recent survey, nearly half of the fresh graduates indicated that they had got at least one part-time job during their college life. Many other investigations also showed that more and more students tend to have a favorable attitude toward part-time jobs. However, there are more parents complaining that these jobs have brought many serious problems like poor academic performance and bad habits. But for my part, I support the view that the students need to find a part-time job for the following three reasons.

In the first place, part-time jobs can increase the students' work and life experience. By doing some part-time jobs, students can enrich their experiences in all sorts of fields. Equipted with these basic and necessary abilities, graduates would find it is easier to find a suitable job. In the second place, it can also bring good changes in students' attitude. Once they have known more about their surroundings, the society and the whole world, they will find that they are more confident and positive than before. The third reason is that study and work can promote each other under certain conditions(A balanced relationship). Rigorous study attitude can be applied to students' work while work experience can also help them expand their thinking.

It looks like you have used the wrong tense of the verb "got" when combined with the helper verb "had".

According to a recent survey, nearly half of the fresh graduates indicated that they had got at least one part-time job during their college life.

Figure 2: Screenshot of error feedback from *WriteToLearn*

To maximize the reliability of scores assigned by human raters, we trained them individually and evaluated their rating performance in a two-stage rating exercise. Initially, a stack of 30 essays (15 essays per prompt) was marked by all raters and rater performance was examined through MFRM analysis. We found that all raters, except Rater 2, achieved satisfactory reliability in marking and an adequate fit to the model (with both the infit and outfit MNSQ values falling between 0.5 and 1.5) as suggested by Linacre (2013a). Rater 2, because of his erratic rating performance (infit MNSQ = 1.59; outfit MNSQ = 1.60), underwent more coaching at the second stage of rating and subsequently managed to reach the expected range of infit and outfit between 0.5 and 1.5.

Table 2: Demographic Information concerning the Human Raters (based on Aryadoust & Liu, 2015)

Raters	**1**	**2**	**3**	**4**
Gender	Female	Male	Male	Female
First Language	English	English	Chinese	Chinese
Highest academic degree	MA	MA	PhD Candidate	MA
Teaching experience	18 years	6 years	6 years	6 years
Rating experience	RCS, RME	RCS	RCS, RME	RCS, RME

Note: RCS= rating in classroom settings; RME= rating in major examinations

Procedures

We chose two preloaded writing prompts from *WriteToLearn* for data collection. Prompt 1 focused on ‘The effect of the Internet on children’ (Expository) and Prompt 2 on 'Whether students should have part-time jobs' (Persuasive). Both prompts were assigned as homework to the participants who were tasked to type out their essays and submit them via email. A total of 326 essays were collected.

Initially, all raters marked all essays using *WriteToLearn*'s analytical scoring rubric. The rubric measured six writing traits: Ideas, Organization, Conventions, Sentence Fluency, Word Choice, and Voice. Each trait was marked on a six-point scale ranging from 1 to 6. Every scale point had thorough descriptors, but due to the confidentiality of the rubric, they cannot be presented.

Then, a batch of 60 essays (30 essays per prompt) were randomly selected for error annotation. All four human raters annotated errors to avoid the potential underestimation of precision and recall (Tetreault & Chodorow, 2008b). As noted earlier, *WriteToLearn* did not categorize errors. Therefore, we developed a set of error codes to facilitate human raters' annotation of errors (adapted from Dikli and Bleyle (2014) and Ferris, Liu, Sinha and Senna (2013) and then mapped the error feedback provided by *WriteToLearn* to the equivalent human raters' error codes. Based on the finally mapping relationships, all raters annotated errors independently and only those that were unanimously coded as a certain type of error were included in the calculation of precision and recall.

Data analysis

Evaluating *WriteToLearn*'s scoring performance

We calibrated *WriteToLearn*'s scoring performance against those of human raters, using MFRM analysis on *FACETS* (Version 3.71; Linacre, 2013b). MFRM offers several important advantages over the classical true score method of establishing rater reliability. First, the classical true score method only provides a broad overview of the instrument or raters; for example, it can estimate whether the scores assigned by raters are in agreement. In contrast, MFRM provides sufficient information for each student writer, rater, scoring traits, and other facets that have been used in the analysis. For example, it estimates the severity level of each rater. Raters whose performances are not satisfactory can be retrained or their scores can be remedied in different ways. Second, MFRM does not require normality of score distributions and, therefore, it can be applied to classroom data where, due to smaller sample sizes, the normality assumption may not be met.

We used a rating scale model and calibrated four facets including student ability, rater severity, writing prompt difficulty and the scoring traits. That is, we postulated that these facets affect the scores that are assigned to the essays. The MFRM analysis can be mathematically expressed as follows:

$$\log \frac{p_{nhijk}}{p_{nhijk-1}} = B_n - C_h - D_i - E_j - F_k$$

where,

P_{nhijk} is the probability of student *n* being awarded a rating of *k* on prompt *i* on trait *j* by rater *h*

$P_{nhijk-1}$ is the probability of student *n* being awarded a rating of *k*–1 on prompt *i* on trait *j* by rater *h*

B_n is the ability of student *n*

C_h is the severity level of rater *h*

D_i is the difficulty of prompt *i*

E_j is the difficulty level of trait $_j$

F_k is the difficulty level of receiving a rating of *k* relative to a rating of *k*–1

Evaluating the accuracy of *WriteToLearn*'s error feedback

Two statistical indices, i.e., precision and recall, were used to evaluate the quality of the feedback provided by *WriteToLearn*. *Precision* measures 'how often the system is correct when it reports that an error has been found' (Leacock *et al.*, 2010: 38). It is the system's hits (i.e., number of cases in which human raters agree that the error identified by *WriteToLearn* was a true error) divided by the total number of errors flagged by the system (Burstein *et al.*, 2003). The minimum threshold of precision for the automated feedback is 90% (Burstein *et al.*, 2003). *Recall* measures 'the system's coverage, i.e., the fraction of errors that the system has detected' (Leacock *et al.*, 2010: 38). It is the number of an AWE system's hits divided by the total number of errors annotated by human raters (Burstein *et al.*, 2003).

Results

WriteToLearn's scoring performance

Table 3 presents descriptive statistics of ratings by four human raters and *WriteToLearn*. Column 1 presents the six traits that were evaluated in each essay; the other five columns display the mean and standard deviation of ratings awarded by four human raters and *WriteToLearn*. On the whole, all human raters rated 326 essays, whereas *WriteToLearn* rated only 319 essays; it failed to rate seven essays. In addition, *WriteToLearn* appeared to be more severe than trained human raters when rating essays, with its mean ratings on the trait of Ideas, Organization, and Voice (M = 2.92, 2.93, and 3.04, respectively) lower than the most severe human rater.

Table 4 presents measurement results of the rater facet through MFRM analysis. Column 1 shows raters' IDs; W*TL* refers to *WriteToLearn* while HR stands for human raters. Column 2 shows the total ratings awarded by each rater and the total number of ratings they have awarded. Overall, each human

Table 3: Descriptive Statistics of Ratings by Four Human Raters and *WriteToLearn*

	HR1		HR2		HR3		HR4		WTL	
	M	SD	M	SD	M	SD	M	SD	M	SD
Ideas	3.77	0.67	4.01	0.92	3.77	0.78	4.05	0.94	2.92	0.59
Organization	3.91	0.45	4.33	0.91	3.95	0.75	3.96	0.90	2.93	0.47
Conventions	4.02	0.32	4.03	1.06	3.66	0.90	3.80	0.85	3.74	0.64
Sentence Fluency	3.52	0.64	4.06	0.99	3.62	0.75	3.92	0.90	3.63	0.65
Word Choice	3.34	0.58	3.87	0.90	3.48	0.76	3.88	0.79	3.40	0.58
Voice	3.80	0.51	3.96	0.98	3.70	0.77	3.75	0.91	3.04	0.63

Note: HR= human Rater; WTL= *WriteToLearn; N*=326 for human raters and *N*=319 for *WriteToLearn*

rater awarded a total of 1,956 ratings while *WriteToLearn* awarded only 1,914 ratings (as *WriteToLearn* failed to score seven essays).

Column 3 presents the severity measures of raters: *WriteToLearn* was most severe (0.95 logits) while HR2 was most lenient (–0.64 logits). The differences of severity between raters were statistically significant, according to the separation coefficient (18.42) and reliability coefficient (1.00) ($\chi^2 = 1{,}322.0$, $df = 162$, $p < 0.01$). That is, the ratings assigned by the raters can be grouped into around 18 levels of difficulty which are significantly different from each other. This suggests that raters have introduced a large amount of variation in the data. Therefore, it does matter whether a paper is marked by human raters or by *WriteToLearn. If* a paper is rated by *WriteToLearn*, it will most likely receive a much lower rating than if it were rated by a human rater.

All raters fit the model reasonably well, although HR1 had a rather low infit and outfit MNSQ index (0.64) which was close to an overfitting pattern. This meant that the rater had chosen a limited range. *WriteToLearn* had the best fit to the model (infit and outfit MNSQ = 1.02 and 1.01). This shows that although its rating algorithm was the most severe of all raters, it rated papers consistently and its consistency is higher than that of the human raters. In addition, the observed exact agreement among raters (37.8%) approximately equaled the expected agreement (37.7%). This indicates that the raters have been behaving like independent experts as expected by the model.

As noted earlier, seven essays were not rated by *WriteToLearn*. These essays were # E29_P1, # E34_ P2, # E109_P1, # E127_P1, # E134_P1, # E151_P1, and # E151_P2. Notably, both # E151_P1 and # E151_P2 were written by Participant # 151, the most misfitting student (Infit MNSQ = 2.48, Outfit MNSQ = 2.46). This indicates that the scores awarded to Participant # 151 have likely been affected by construct-irrelevant factors. To investigate the possible causes, we conducted a post hoc analysis of the two essays written by Participant # 151. This is discussed in the Discussion section.

Table 4: Measurement Results of Raters

	Total				Infit		Outfit	
Rater	score	count	Severity	SE	MNSQ	ZSTD	MNSQ	ZSTD
WTL	6272	1914	0.95	0.03	1.02	0.51	1.01	0.38
HR3	7230	1956	0.05	0.03	0.92	−2.50	0.93	−2.31
HR1	7286	1956	−0.01	0.03	0.64	−9.00	0.64	−9.00
HR4	7614	1956	−0.35	0.03	1.09	2.61	1.09	2.81
HR2	7909	1956	−0.64	0.03	1.29	8.28	1.29	8.29

Separation: 18.42; Reliability: 1.00
Fixed (all same) chi-square (χ^2): 1322.0, $df = 162$, $p < 0.01$
Exact agreement: 37.8%; expected: 37.7%

Note: *WTL*= *WriteToLearn*; HR= human rater

WriteToLearn's error feedback accuracy

Table 5 presents the results of the accuracy analysis of *WriteToLearn's* feedback. The two far left columns demonstrate the 22 error types and the total number of errors of each error type identified by human raters in the subsample of 60 essays. Among the 22 error types, the three most frequent errors were word choice (n = 125), capitalization (n = 115), and article errors (n = 115). Students also made a lot of singular/plural (n = 86), preposition (n = 79), expression (n = 78), and pronoun (n = 60) errors.

The four columns under *WriteToLearn* display the results of the evaluation of its error detection ability. Column 1 under *WriteToLearn* shows the total number of each error type identified by *WriteToLearn*. The system detected 15 out of 22 error types identified by human raters and missed the other seven types of errors, including expression, modal verbs, passive voice, sentence structure, verb tense, word form, and word order. Among the 15 error types it did identify, *WriteToLearn* was more capable of detecting capitalization (n = 104), spelling (n = 93) and punctuation (n = 92) errors; whereas other error types were largely ignored by it (with the total number of detected errors ranging from 1 to 19). The following example illustrates the comma errors correctly detected by *WriteToLearn*:

(1) The Internet also bring a lot of pleasure to children $\overset{\wedge}{\underline{\quad}}$[1] which can let children feel relaxed. (# E16_P1)

WriteToLearn's feedback: Comma may be missing before or after a nonrestrictive element. 'Which' is used to introduce a word, phrase, or clause that is not essential to the meaning of the sentence. Set the element apart using a comma, or use 'that' instead of 'which'.

Column 2 under *WriteToLearn* reports its hits. Except for capitalization, punctuation, spelling and subject-verb agreement, few errors of other error types were correctly identified by the system. Additionally, five error types were totally misidentified, including fragment, possessive case, preposition, verb form and word choice, many of which were frequent errors committed by ELL students (see Leacock *et al*., 2010). For example, word choice is the most salient error type in this study both in terms of its high frequency and *WriteToLearn's* failure to detect such errors. *WriteToLearn* identified seven errors against human raters' 125 word-choice errors but none of them were correctly identified. Example 2 illustrates a word choice error that was falsely flagged:

(2) It is conspicuous that the effects not only exist among adults, but also exist among children. (# E80_ P1)

WriteToLearn's feedback: Considering using 'affects' instead of 'effects' here.

Additionally, *WriteToLearn* also missed a lot of word choice errors detected by human raters, shown in Examples 3 and 4 as follows:

(3) First and foremost, having jobs while studying is an effective way to adopt ourselves to social life. (# E24_P2)

(4) For example, there are some sexy news and pictures on web which will not do good to teenagers. (# E40_P1)

In Example 3, the student should have used the word 'adapt' instead of 'adopt'. In Example 4, what the student intended to express was that there was much information about sex on the Internet, which would be harmful to teenagers.[2] Such word choice errors were frequently found in this study; unfortunately, *WriteToLearn* did not identify any of these errors.

Closely related with word choice errors, problematic expressions including both non-idiomatic and unclear expressions (that are difficult for human raters to decipher the meaning of) were frequently found. In total, human raters annotated 78 such expression errors, none of which were detected by *WriteToLearn*. Examples 5 and 6 show two sample errors of expression:

(5) Nowadays, there is a prevalent phenomenon that some teenagers spend money like water. (# E130_ P2)

(6) It is advisable for us to tell children how to discard the dregs while keeping the essence as much as possible. (# E12_P1)

Both of these are problematic English expressions that are translations of old Chinese sayings word-by-word. The intended meaning of Example 5 is that 'Nowadays, teenager would always use up all their money' while Example

Table 5: Human Rater's and *WriteToLearn*'s Error Feedback Results

Error type	**Human rater's**	***WriteToLearn's***			
	Total	Total	Hits	Precision (%)	Recall (%)
Connecting words	18	1	1	100	5.6
Capitalization	115	104	96	92.3	83.5
Subject-verb agreement	42	19	15	79	35.7
Comma splice	10	8	6	75	60
Singular/plural	86	12	9	75	10.5
Article	115	10	7	70	6.1
Run-on sentences	8	14	6	42.9	75
Punctuation	54	92	34	37	63
Spelling	52	93	18	19.4	34.7
Pronoun	60	7	1	14.3	1.7
Fragment	13	5	0	0	0
Possessive case	4	6	0	0	0
Preposition	79	1	0	0	0
Verb form	33	5	0	0	0
Word Choice	125	7	0	0	0
Expression	78	0	0	N/A	0
Modal verbs	10	0	0	N/A	0
Passive voice	12	0	0	N/A	0
Sentence structure	47	0	0	N/A	0
Verb tense	15	0	0	N/A	0
Word form	41	0	0	N/A	0
Word order	15	0	0	N/A	0
Total	1032	394	193	49	18.7

Note: Precision = Hits divided by *WriteToLearn*'s total (For example, the precision of capitalization: 96÷104 = 92.3); Recall = Hits divided by human rater's total (For example, the recall of capitalization: 96 ÷115 = 83.5).

6 actually means 'It is advisable for us to tell children how to make good use of the advantages of the Internet and avoid being influenced by its negative effects'.

Columns 3 and 4 under *WriteToLearn* present the precision and recall of its error feedback. *WriteToLearn* was most precise in detecting connecting words errors, as indicated by its precision of 100%. However, its recall was only 5.6%, indicating that it missed 94.4% of the errors identified by human raters. It was also highly precise in detecting capitalization errors, with a precision rate of 92.3% and recall of 83.5%. It demonstrated a similar pattern in identifying subject-verb agreement, comma splice, and singular/plural nouns and articles errors with precision rates varying between 70% and 79% and recall rates below 50% (the only exception was comma splice with its recall of 60%). In its identification of run-on sentences, punctuation, spelling and pronoun errors, *WriteToLearn* demonstrated less reliable performance with precision rates below 50% and recall rates ranging from 1.7% to 75%. Its precision and recall rates of fragment, possessive case, preposition, verb form and word choice errors were 0% because it misidentified all of these five types of error (as shown in the column of Hits). *WriteToLearn*'s overall precision and recall in detecting errors out of the 22 error types were 49% and 18.7% respectively, meaning that it missed 81.3% errors that were identified by human raters and only 49% of the errors detected were true errors.

Discussion

WriteToLearn's scoring ability

Reliability and severity

The main positive finding from this study regarding *WriteToLearn's* scoring ability is that it had the best fit to MFRM, suggesting that it consistently marked all essays and that its consistency was higher than that of human raters. Therefore, teachers who plan to use *WriteToLearn* in the classroom can be assured of its reliability.

However, a word of caution is in order. The results also showed that there were marked differences in severity between raters, with *WriteToLearn* being the most severe rater. Therefore, we would recommend that when *WriteToLearn* is used to assess students' writing in a classroom setting, its automated scores should be adjusted according to the different level of rater severity; otherwise, the significant discrepancy of severity between *WriteToLearn* and human raters would jeopardize the validity of score-interpretations in terms of the estimation of students' writing ability. To ascertain the fairness of the scores, teachers can randomly check the automated scores and decide whether students have been under-rated, or to provide their own scores along with the automated scores generated by *WriteToLearn*.

Off-topic essays

As noted earlier, *WriteToLearn* failed to score seven essays. We conducted a post hoc analysis of these essays in order to investigate the possible reasons for *WriteToLearn's* failure to score them. According to human raters, two essays written by Participant # 151 (i.e., # E151_P1 and # E151_P2) are of different qualities. All four human raters rated # E151_P1 as off-topic because it did not discuss the effects of the Internet on children as required by the prompt. Consequently, they penalized the trait score of Ideas. Despite the erratic performance in the first essay, Participant # 151 performed well in the persuasive essay based on prompt 2. All human raters awarded medium or high scores to all of the six traits of the essay: (HR1) 4, 5, 5, 4, 5, 5, (HR2) 4, 4, 3, 3, 5, 4, (HR3) 4, 3, 4, 4, 4, 3, and (HR4) 5, 4, 4, 3, 4, 4.

While the human raters identified the differences in quality between the two essays, *WriteToLearn* made no distinction between them. It failed to detect the off-topicness of # E151_P1 and to recognize the high quality of # E151_P2. In the scoreboards for both essays, *WriteToLearn* provided no score for the six traits but a general comment 'Please review your essay with your teacher. Your choice of topic, style, vocabulary, or organization is different than others written in response to this prompt.'

Based on these examples, it could be concluded that *WriteToLearn* seemed to follow 'a typical methodology' across all current AWE systems (McNamara *et al*., 2015: 37). They do not have the capability to 'read' and parse essays to evaluate their quality as human raters do; instead, they evaluate essays based on linguistic features that have been extracted from essays on which they have been trained. Therefore, they are unable to score essays that do not include those linguistic features. This conclusion was confirmed by the explanation given by *WriteToLearn*'s developer after we reported the failure to score several essays: Before scoring a response, the scoring engine first evaluates the accuracy with which it can do so. If the confidence value is too low, the student will receive a message that the response cannot be scored. *The confidence value will be low if the student's essay does not look enough like the many essays on which the scoring engine has been trained* (Pearson Education Inc., personal communication, 11 April 2014; emphasis added).

Validity and usefulness

WriteToLearn's failure to read essay content and to make discerning judgments about students' writing resonates with previous research on *WriteTo Learn* (McGee, 2006), *My Access* (Hoang, 2011), and *Criterion* (Powers, 2000). These findings provide justification for the concern over the validity of automated scores (Attali, 2013; Weigle, 2013b) and pose a challenge to the use of such scores as independent indicators of students' writing ability particularly in classroom settings.

Thus, at least two general concerns remain. First, the measures that AWE systems employ (i.e., a set of linguistic features with corresponding weightings) may not fully represent the real-world construct of writing which are more than linguistic features. Therefore, when an AWE system has difficulty with rating essays or is too severe, the validity of the scores are questionable. Further, any revisions made by students based on the AWE scores may only be targeting many linguistic features. And, an increase in automated scores after any revisions may not necessarily indicate improvement in writing. This point has also asserted by researchers including Stevenson and Phakiti (2014). Second, the use of automated scores as criterion measures may lead to both teachers' and students' over-dependence on the score increase across drafts rather than on improvement in writing. This point has also been asserted by Warschauer and Grimes (2008).

In conclusion, given the consistency of *WriteToLearn* along with the concerns mentioned above, it is suggested that *WriteToLearn* be implemented as 'partial summative assessment tools' (Li *et al.*: 76) and its automated scores not be used as the sole indicator of student writing ability but be used along with teacher evaluations of students' writing.

***WriteToLearn*'s error detecting ability**

Precision and recall

The results from the study showed that *WriteToLearn's* overall error detecting performance (precision = 49%; recall = 18.7%) did not meet the threshold of 90% precision suggested by Burstein *et al.* (2003). Notably, *WriteToLearn* misidentified or failed to detect error types that stand out as particularly salient in the writing samples of the present study including word choice, article and preposition errors. These findings are consistent with previous findings of AWE systems' poor performance in detecting errors made by ELL students (see Dikli & Bleye, 2014; Hoang, 2011). This lends further credit to the argument that such errors may pose particular difficulty to current AWE systems which were not initially designed for the ELL writers but have been gradually marketed to schools and colleges where writers are ELLs.

Use in classrooms

WriteToLearn's unreliable error detection may render its efficacy for use in ELL instructional settings questionable. First, students may be confused by potential erroneous automated feedback. They may make inaccurate amendments to the text (see Galleta, Durcikova, Everard, & Jones, 2005), which may impede their development of writing proficiency. Second, the low accuracy of automated feedback may lead to students' negative perception of or resistance to the implementation of AWE systems (see Chen & Cheng, 2008). Low uptake of automated feedback in their revision process is a clear example of this problem (see Attali, 2004; Li, Link, & Hegelheimer, 2015).

However, *WriteToLearn*'s difficulty in detecting certain error types should be interpreted with caution. Research has shown that the manner in which AWE systems is integrated into writing instruction influences teacher and student perceptions and how students use it (Chen & Cheng, 2008; Li *et al.*, 2015). Weigle (2013a) argues that such systems can be put to 'beneficial use when implemented with forethought and understanding of their constraints and limitations' (p. 97).

Understanding strengths and limitations

One of the main attractions of automated feedback like *WriteToLearn* is that it is instantaneous and readily available. This immediacy can motivate students to write and produce multiple drafts on the same prompts, hence improving their grammatical accuracy and increasing their learning autonomy (see Dikli & Bleyle, 2014; Foltz *et al.*, 1999). The use of automated feedback may also liberate writing teachers from the heavy burden of providing timely feedback on students' initial drafts. This may also enable them to focus on higher-level concerns such as argumentation, organization, and voice in students' writing (see Chen & Cheng, 2008; Li *et al.*, 2015; Warschauer & Grimes, 2008).

But for automated feedback to be integrated effectively into ELL writing classrooms, teachers and students should be aware of the strengths and limitations. Teachers should caution students that some of the feedback they receive might not be necessarily helpful and encourage students to adjust their uptake of automated feedback on different error types according to their accuracy. Additionally, teachers would need to provide their own feedback on the errors that are frequently misidentified or unidentified instead of leaving students to merely rely on automated feedback to revise their essays.

Conclusion

This study investigated the application of *WriteToLearn* to Chinese undergraduate English majors by evaluating its scoring performance and accuracy of error feedback as compared to trained human raters. The findings from the present study showed that its strength lies in its consistency in scoring and timely feedback although it also was the most severe when compared to four trained human raters and the feedback was not always accurate and useful. But it was *WriteToLearn's* difficulty in identifying off-topic essays correctly and providing accurate and useful diagnostic feedback that was problematic.

The findings also provided some ideas for the application of *WriteToLearn* in ELL classroom settings. First, teachers should always have a clear understanding of *WriteToLearn's* strengths and limitations. They can then help students recognize and compensate for such limitations. Second, *WriteToLearn* has the potential to act as a time-saving and cost-effective instructional tool in ELL writing classrooms. For example in China, large class sizes make it difficult for teachers to score and provide timely feedback on students' essays.

Therefore, *WriteToLearn* should be given an appropriate role in ELL writing instruction: when it is to be used for assessing student writing, it could be used as a second or a third rater along with one or more trained human raters; the combination of automated feedback from *WriteToLearn* and teacher and peer feedback may be beneficial to students' writing improvement.

The findings also have an important methodological implication for AWE research. It showed the feasibility of MFRM analysis in evaluating AWE systems' scoring ability. In addition to scoring consistency, MFRM analysis can provide robust analyses of AWE systems' rating severity and the functionality of their preloaded writing prompts and scoring rubrics. MFRM can also be used to capture rater bias patterns, i.e., whether a rater has been substantially biased against test-takers with certain backgrounds (e.g., male vs female), tasks (e.g., expository vs persuasive writing), and scoring traits in analytical scoring rubrics.

Acknowledgement

This research was supported by the Fundamental Research Funds of China West Normal University.

Notes

1. The error type under discussion in the example sentences are underlined for emphasis throughout this paper.
2. The writers of Example 4, 5 and 6 were involved in the interpretation.

About the authors

Sha Liu is assistant lecturer at School of Foreign Languages at China West Normal University in People's Republic of China. She teaches English Essay Writing and Integrated English Course to English majors. Her research focuses on second language writing assessment and the application of automated writing evaluation to classroom settings.

Antony John Kunnan is Professor of English Language at Nanyang Technological University, Singapore. He has published widely in the area of language assessment, especially, on validation, test bias, and language assessment policy. His recent publications include a four-volume edited collection of original chapters titled *The Companion to Language Assessment* (Wiley, 2014) and a four-volume edited collection of published papers titled *Language Testing and Assessment* (Routledge, 2015). He was the founding editor of *Language Assessment Quarterly* (2003–2013), past president of the International Language Testing Association and current president of the Asian Association for Language Assessment.

References

Aryadoust, V., & Liu, S. (2015). Predicting EFL writing ability from levels of mental representation measured by Coh-Metrix: A structural equation modeling study. *Assessing Writing,* 24, 35–58. http://dx.doi.org/10.1016/j.asw.2015.03.001

Attali, Y. (2013). Validity and reliability of automated essay scoring. In M. D. Shermis & J. Burstein (Eds), *Handbook of automated essay evaluation: Current applications and new directions,* 181–199. New York, NY: Routledge.

Attali, Y., & Burstein, J. (2005). Automated essay scoring with e-rater® V.2.0 (ETS research report number RR-04-45). Retrieved from http://www.ets.org/Media/Research/pdf/RR-04-45.pdf

Bridgeman, B. (2013). Human ratings and automated essay evaluation. In M. D. Shermis & J. Burstein (Eds), *Handbook of automated essay evaluation: Current applications and new directions,* 221–232. New York: Routledge.

Burstein, J. (2003). The e-rater® scoring engine: Automated essay scoring with natural language processing. In M. D. Shermis & J. Burstein (Eds), *Automated essay scoring: A cross-disciplinary perspective,* 113–121. Mahwah, NJ: Lawrence Erlbaum Associates.

Burstein, J., Chodorow, M., & Leacock, C. (2003). Criterion online essay evaluation: An application for automated evaluation of student essays. *Proceedings of the Fifteenth Annual Conference on Innovative Applications of Artificial Intelligence*, Acapulco: Mexico.

Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M., Braden-Harder, L., & Harris, M. D. (1998, August). Automated scoring using a hybrid feature identification technique. *Proceedings of the Annual Meeting of the Association of Computational Linguistics*, Montreal. Retrieved from http://www.ets.org/Media/Research/pdf/erater_acl98.pdf

Chen, C. F., & Cheng, W. Y. (2008). Beyond the design of automated writing evaluation: Pedagogical practices and perceived learning effectiveness in EFL writing classes. *Language Learning & Technology,* 12 (2), 94–112.

Dikli, S., & Bleyle, S. (2014). Automated Essay Scoring feedback for second language writers: How does it compare to instructor feedback? *Assessing Writing,* 22, 1–17. http://dx.doi.org/10.1016/j.asw.2014.03.006

Ferris, D. R., Liu, H., Sinha, A., & Senna, M. (2013). Written corrective feedback for individual L2 Writers. *Journal of Second Language Writing,* 22, 307–329. http://dx.doi.org/10.1016/j.jslw.2012.09.009

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Educational Journal of Computer-Enhanced Learning,* 1 (2). Retrieved from http://imej.wfu.edu/articles/1999/2/04/printver.asp

Foltz, P. W., Lochbaum, K. E., & Rosenstein, M. R. (2011, April). *Analysis of student ELA writing performance for a large scale implementation of formative assessment*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, Louisiana.

Foltz, P. W., Streeter, L. A., Lochbaum, K. E., & Landauer, T. (2013). Implementation and

Application of the Intelligent Essay Assessor. In M. D. Shermis & J. Burstein (Eds), *Handbook of automated essay evaluation: Current applications and new directions,* 66–88. New York, NY: Routledge.

Galleta, D. F., Durcikova, A., Everard, A., & Jones, B. (2005). Does spell-checking software need a warning label? *Communication of the ACM*, 48 (7), 82–85. http://dx.doi.org/10.1145/1070838.1070841

Han, N., Chodorow, M., & Leacock, C. (2006). Detecting errors in English articles usage by non-native speakers. *Natural Language Engineering,* 12 (2): 115–129. http://dx.doi.org/10.1017/S1351324906004190

Hoang, G. (2011). *Validating My Access as an automated writing instructional tool for English language learners* (Unpublished Master's thesis). California State University, Los Angeles.

Koskey, K., & Shermis, M. D. (2013). Scaling and norming for automated essay scoring. In M. D. Shermis & J. Burstein (Eds), *Handbook of automated essay evaluation: Current applications and new directions,* 200–220. New York: Routledge.

Landauer, T. K., Laham, D., & Foltz, P. W. (2003). Automatic essay assessment. *Assessment in Education,* 10, 295–308. http://dx.doi.org/10.1080/0969594032000148154

Leacock, C., Chodorow, M., Gamon, M., & Tetreault, J. (2010). Automated grammatical error detection for language learners. *Synthesis Lectures on Human Language Technologies,* 3, 1–34. http://dx.doi.org/10.2200/S00275ED1V01Y201006HLT009

Li, J., Link, S., & Hegelheimer, V. (2015). Rethinking the role of automated writing evaluation (AWE) feedback in ESL writing instruction. *Journal of Second Language Writing,* 27, 1–18. http://dx.doi.org/10.1016/j.jslw.2014.10.004

Li, Z., Link, S., Ma, H., Yang, H., & Hegelheimer, V. (2014). The role of automated writing evaluation holistic scores in the ESL classroom. *System,* 44, 66–78. http://dx.doi.org/10.1016/j.system.2014.02.007

Linacre, J. M. (2013a). *A user guide to Facets, Rasch-model computer programs*. Chicago, IL: Winsteps.com.

Linacre, J. M. (2013b). Facets Rasch measurement [computer program]. Chicago, IL: Winsteps.com.

McGee, T. (2006). Taking a spin on the Intelligent Essay Assessor. In P. F. Ericsson & R. H. Haswell (Eds), *Machine scoring of student essays: Truth and consequences,* 79–92. Logan, UT: Utah State University Press.

McNamara, D. S., Crossley, S. A., Roscoe, R. D., Allen, L. K., & Dai, J. (2015). A hierarchical classification approach to automated essay scoring. *Assessing Writing,* 23, 35–39. http://dx.doi.org/10.1016/j.asw.2014.09.002

Pearson Education Inc. (2010). *Intelligent Essay Assessor (IEA) fact sheet*. Retrieved from http://kt.pearsonassessments.com/download/IEA-FactSheet-20100401.pdf

Perelman, L. (2014). When 'the state of the art' is counting words. *Assessing Writing,* 21, 104–111. http://dx.doi.org/10.1016/j.asw.2014.05.001

Powers, D. E. (2000). *Computing reader agreement for the GRE Writing Assessment* (ETS research memorandum, RM-00-08). Princeton, NJ: Educational Testing Service.

Powers, D. E., Burstein, J., C., Chodorow, M., Fowels, M. E., & Kukich, K. (2002). Stumping e-rater: Challenging the validity of automated essay scoring. *Computers in Human Behavior,* 18 (2), 103–134. http://dx.doi.org/10.1016/S0747-5632(01)00052-8

Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration, *Assessing Writing,* 20, 53–76. http://dx.doi.org/10.1016/j.asw.2013.04.001

Shermis, M. D., & Burstein, J. C. (2003). Introduction. In M. D. Shermis & J. Burstein (Eds), *Automated essay scoring: A cross-disciplinary perspective,* xiii–xvi. Mahwah, NJ: Lawrence Erlbaum Associates.

Stevenson, M., & Phakiti, A. (2014). The effects of computer-generated feedback on the quality of writing. *Assessing Writing,* 19, 51–65. http://dx.doi.org/10.1016/j.asw.2013.11.007

Tetreault, J., & Chodorow, M. (2008a, August). *The ups and downs of preposition errors detection in ESL writing*. Proceedings of the 22nd International Conference on Computational Linguistics (COLING), Manchester, UK. http://dx.doi.org/10.3115/1599081.1599190

Tetreault, J., & Chodorow, M. (2008b, August). *Native judgments of non-native usage: Experiments in preposition error detection*. Proceedings of the Workshop on Human Judgments in Computational Linguistics at the 22nd International Conference on Computational Linguistics (COLING), Manchester, UK. http://dx.doi.org/10.3115/1611628.1611633

Vantage Learning. (2003a). *Assessing the accuracy of Intellimetric for scoring a district-wide writing assessment (RB-806)*. Newton, PA: Vantage Learning.

Vantage Learning. (2003b). *How does Intellimetric score essay response?* (RB-929). Newton, PA: Vantage Learning.

Vantage Learning. (2006). *Research summary: Intellimetric scoring accuracy across genres and grade levels*. Retrieved from http://www.vantagelearning.com/docs/intellimetric/IM_ReseachSummary_InteliMetric_Accuracy_Across_Genre_and_Grade_Levels.pdf

Warschauer, M., & Grimes, D. (2008). Automated writing in the classroom. *Pedagogies: An International Journal,* 3 (1), 22–26. http://dx.doi.org/10.1080/15544800701771580

Warschauer, M., & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language teaching research,* 10 (2), 157–180. http://dx.doi.org/10.1191/1362168806lr190oa

Weigle, S. C. (2013a). English language learners and automated scoring of essays: Critical considerations. *Assessing Writing,* 18, 85–99. http://dx.doi.org/10.1016/j.asw.2012.10.006

Weigle, S. C. (2013b). English as a second language writing and automated essay evaluation. In M. D. Shermis & J. Burstein (Eds), *Handbook of automated essay evaluations: Current applications and new directions,* 36–54. New York: Routledge.